Speech Emotion Recognition Using Non-Linear Teager Energy Based Features in Noisy Environments

Alexandros Georgogiannis, Vassilis Digalakis
Department of Electrical and Computer Engineering
Technical University of Crete, Chania 73100, Greece

***Abstract*—In this study, Teager-energy based Mel-frequency cepstral coefficients (TEMFCCs) are proposed for Automatic Speech Emotion Recognition (ASER) in noisy environments. TEMFCCs are obtained by taking the absolute value of the Teager-energy operator (TEO) of the short-time Fourier transform of the signal (STFT), warping it to a Mel-frequency scale, and taking the discrete cosine transform (DCT) of the log-Mel Teager-energy spectrum. Experiments on classification of discrete emotion categories show that TEMFCCs are more robust than MFCCs in noisy conditions, while TEMFCCs and MFCCs perform similarly for clean conditions.**

***Index Terms*—emotion recognition, speech analysis, nonlinear acoustics**

I. Introduction

Automatic speech emotion recognition (ASER) refers to the task of classifying speech phrases into emotional classes. Although it is a relatively new field of research, it has many potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive to their emotions.

Many techniques for employing ASER are drawn from the field of automatic speech recognition (ASR). Despite recent advances in the state-of-the art of ASER, ASER in noisy conditions remains an open research problem. The techniques that have been proposed in the literature for improving the robustness of speech emotion recognition in noise mainly fall into three categories: acoustic model-adaptation algorithms, speech-enhancement algorithms and robust feature-extraction algorithms. We concentrate in the problem of robust feature extraction. The TEMFCC feature set we propose is motivated by the nonlinear TEO operator that estimates the energy of the source of a resonance signal, and MFCCs.

MFCCs are one of the most widely used cepstral representations for ASER but are easily affected by common frequency-localized random perturbations, to which human perception is largely insensitive [4]. MFCCs' degradation of performance led researchers to utilize TEO in the development of new cepstral representations. The effect of noise can be eliminated by using the TEO in feature extraction [11]. Teager-energy based mel-frequency cepstrum coefficients (T-MFCCs) [9] were developed for language identification (LID) and make use of the TEO of the signal. In this paper we design a front-end that combines a mel-filterbank with the Teager-energy estimation method. The proposed features are evaluated on speech emotion recognition tasks in noise and are shown to be more robust than the MFCCs and T-MFCCs.

II. Teager Energy Operator

Newton's law of motion for an oscillator with mass m and spring constant k states that

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \tag{1}$$

and its solution consists of a signal $x(t) = acos(\phi(t))$. The system's total energy E is the sum of the kinetic and potential energy and is given by

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \Rightarrow E = \frac{1}{2}m\omega^2 a \tag{2}$$

where $\omega = d\phi(t)/dt$.

Motivated by the above analysis of the energy of an oscillating system, Teager and then Kaiser [5] proposed the need for time-frequency analysis methods that can track rapid energy changes within a glottal cycle. This led to the definition of TEO based on a definition of energy that accounts for the energy in the system that generated the signal [11].

For the case of real continuous-time signals, TEO is defined as:

$$\Psi(x(t)) = \dot{x}(t)^2 - x(t)\ddot{x}(t) \tag{3}$$

and for real discrete-time signals as:

$$\Psi(x[n]) = x[n]^2 - x[n-1]x[n+1] \tag{4}$$

As an energy operator, we expect TEO to have positive values, but this is not always the case for all signals. For complex continuous-time signals [6], the definition of TEO is:

$$\Phi(x(t)) = \Psi(\Re\{x(t)\}) + \Psi(\Im\{x(t)\}) \tag{5}$$

and for complex discrete-time signals is:

$$\Phi(x[n]) = \Psi(\Re\{x[n]\}) + \Psi(\Im\{x[n]\}) \tag{6}$$

The above definition of energy for complex signals exhibits the symmetry of the operator more clearly. TEO of a complex signal is the sum of the energy of the real and imaginary parts of the signal. Note that both definitions yield a real quantity, as expected for an energy operator.

A. Prior work on TEO in Time-Frequency analysis and Signal Representations

Time-frequency distributions estimating the signal energy content in time and frequency bins are considered indispensable for the study of non-stationary signals, such as speech, radar, geophysical, and biological signals [3]. TEO is frequently used in the development of feature representations based on non-linear transformations.

Thus, the computation of such time-frequency estimations can be generalized as an energy estimation problem in the presence of noise. The most widely used energy estimation scheme is based on the squared energy operator (SEO), $S(\cdot)$, where the squared signal is the instantaneous energy term. The definition of SEO for continuous time signals is:

$$S(x(t)) = x^2(t) \tag{7}$$

and for discrete time signals as:

$$S(x[n]) = x^2[n] \ . \tag{8}$$

Many researchers, in order to gain advantage from the fact that TEO incorporates both amplitude and frequency information, have developed frequency representations that utilize TEO. Our work demonstrates that TEO in frequency domain provides better representations of non linear variations of energy than in time signals. In [8], a system is developed for the detection of human stress and emotions based on TEO and log-frequency power cepstral coefficients (LFPCs). In [4], Teager-energy cepstrum coefficients (TECCs) are proposed that use TEO and a constant Q-gammatone filter-bank. In [7] Teager-energy features derived from the power spectrum difference (PSD) and TEO have been proposed to improve the robustness of speech recognizer in the presence of white noise. In [10], an extended version of TEO is developed called, variable length TEO (VTEO)

$$VTEO_j(x[n]) = x[n]^2 - x[n-j]x[n+j],$$

in order to identify speakers from their "hum"; note that ordinary TEO is a special case for $VTEO_j$ with $j = 1$.

The general process when TEO is applied in the frequency domain is to pass the frame spectrum, $S(k)$, of the signal $s(n)$ through a filter bank (e.g., Mel-scale or Q-gammatone) and then apply TEO in the frequency domain as follows:

$$\Psi(|S_i^m(k)|) = |S_i^m(k)|^2 - |S_i^m(k+1)||S_i^m(k-1)|$$

where $S_i^m(k)$ is the sampled frequency domain output of the i^{th} filter in the m^{th} frame. Instead of using the above approach, we apply TEO on the STFT of signal and thus we make use of the complex version of TEO (equation (6)):

$$\Phi(S_i^m(k)) = \Psi(\Re\{S_i^m(k)\}) + \Psi(\Im\{S_i^m(k)\}) \ .$$

The average energy E_i^m of the i^{th} filter $S_i^m(k)$ in the m^{th} frame is

$$E_i^m = \frac{1}{N_i}\sum_{k=1}^{N_i}|S_i^m(k)|^2, \quad i = 1,\ldots,L, \quad m = 1,\ldots,M \tag{9}$$

where L is the total number of filters in a Mel-scaled filter bank and N_i is the number of frequency coefficients in the i^{th} Mel filter. The average frame energy E_{avg}^m of m^{th} frame is

$$E_{avg}^m = \frac{1}{L}\sum_{i=1}^{L}E_i^m \ . \tag{10}$$

The average Teager-energy T_i^m of the i^{th} Mel filter in the m^{th} frame is

$$T_i^m = \frac{1}{N_i}\sum|\Psi(|S_i^m(k)|)| \tag{11}$$

and the average Teager-energy T_{avg}^m of the m^{th} frame is

$$T_{avg}^m = \frac{1}{L}\sum_{i=1}^{L}T_i^m \ . \tag{12}$$

Similarly, we define the average Teager-energy TC_i^m of the i^{th} mel-filter in the frame of the complex signal $S_i^m(k)$ as

$$TC_i^m = \frac{1}{N_i}\sum|\Phi(S_i^m(k))| \tag{13}$$

and the average Teager-energy TC_{avg}^m of m^{th} frame as

$$TC_{avg}^m = \frac{1}{L}\sum_{i=1}^{L}TC_i^m \ . \tag{14}$$

To show the effectiveness of the features obtained by conventional energy and the Teager-energy, we compared the envelopes of E_{avg}^m, T_{avg}^m, and TC_{avg}^m (for $m = 1,\ldots,M$) of clean and noisy (0 dB white and pink noise) samples of an emotional phrase (anger) uttered by a male speaker (Figure 1). We calculate the root square error (RMSE) between $(E_{avg}^m)_{noisy}$ and $(E_{avg}^m)_{clean}$,

$$E_{RMSE} = \sqrt{\frac{\sum_{m=1}^{M}[(E_{avg}^m)_{noisy} - (E_{avg}^m)_{clean}]^2}{\sum_{m=1}^{M}[(E_{avg}^m)_{clean}]^2}} \tag{15}$$

as a statistical measure of comparison. The above statistical measure is proportional to $1/SNR$, so the lower it is, the better the estimation of energy is. Similarly, we compute the quantities T_{RMSE} and TC_{RMSE} for T_{avg}^m and TC_{avg}^m. Table I shows the calculated RMSE values. Features extracted from T_i^m and TC_i^m give better performance than E_i^m in noisy environment. Hence, the application of the complex TEO in the frequency domain outperforms both the application of the real TEO in the frequency domain and the squared energy estimator as an energy estimator in noisy environments.

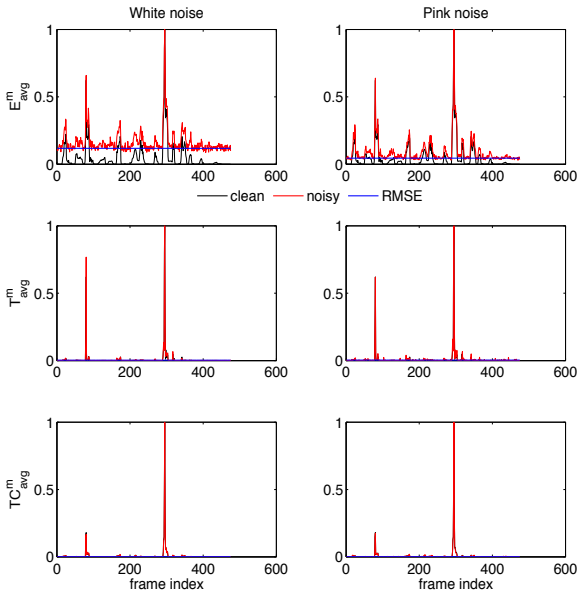


Fig. 1: E^m_{avg}, T^m_{avg}, and TC^m_{avg} envelops for an emotional phrase expressing anger uttered by a male speaker. The left column is for AWGN and the right column for pink noise

III. TEMFCC PARAMETERS

Calculation of TEMFCC requires the following steps:

1) The speech signal $s[n]$ is first passed through a pre-processing stage, which includes frame blocking, hamming windowing with an analysis window $w[n]$, and pre-emphasis, to give the pre-processed speech signal $s[\hat{n}]$.
2) The discrete Fourier transform (DFT), $S[\hat{n}, \omega_k]$, of $s[\hat{n}]$ is computed:

$$S[\hat{n}, \omega_k] = \sum_{m=-\infty}^{\infty} s[m] w[\hat{n} - m] e^{-j\omega_k m} \tag{16}$$

where $\omega_k = \frac{2\pi}{N} k$ and N is the DFT length.
3) The TEO of $S[\hat{n}, \omega_k]$, $\Phi(S[\hat{n}, \omega_k])$ is calculated and the magnitude of $\Phi(S[\hat{n}, \omega_k])$ is then weighted by a mel-filterbank $V_l[\omega_k]$. This filter bank is composed by a series of filter frequency responses whose center frequencies and bandwidths roughly match those of the auditory critical band filters (mel-frequency warping). An example of such a filter is shown below.

TABLE I
RMSE VALUES FOR E^m_{avg}, T^m_{avg}, AND TC^m_{avg}

	White noise	Pink noise
E_{RMSE}	1.0920	0.0408
T_{RMSE}	0.1650	0.0233
TC_{RMSE}	0.0066	0.0076

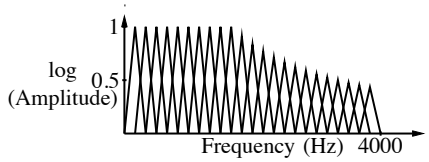


Fig. 2: Triangular mel-scale filter bank with 24 filters

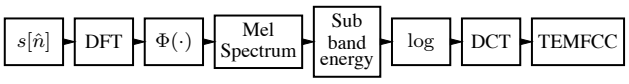
4) We then compute the energy in $\Phi(S[\hat{n}, \omega_k])$, weighted by each mel-scale frequency response. The resulting energies are given for each speech frame at time n and for the l^{th} mel-scale filter, $V_l[\omega_k]$ and $l = 1, \ldots, L$, as

$$e[\hat{n}, l] = \sum_{k=L_l}^{U_l} |V_l[\omega_k] \Phi(S[\hat{n}, \omega_k])| \tag{17}$$

where L is the total number of filters, and L_l, U_l denote the lower and upper frequency indices respectively over which each filter is non-zero.
5) At the last step, the DCT of the log magnitude of the filter outputs for each frame is computed to form TEMFCC$[\hat{n}, \omega_k]$, i.e.,

$$\underset{k=1,\ldots,N_c}{\text{TEMFCC}[\hat{n}, \omega_k]} = \frac{1}{L} \sum_{l=1}^{L} \log(e[\hat{n}, l]) \cos\left(\frac{k(l - 0.5)}{L}\right) \tag{18}$$

Figure below is a flow-diagram describing the computation of TEMFCCs. The proposed method is also compared with the MFCC feature parameters and the T-MFCC features parameters. The T-MFCCs features employ TEO in the time domain (compared to TEMFCCs that employ TEO in the frequency domain). Next we briefly explain T-MFCC.



A. T-MFCC parameters

T-MFCCs feature parameters were developed for language-identification (LID). LID refers to the task of identifying an unknown language from the test utterances. The computation of T-MFCCs for a signal $s[n]$ requires the following steps:

1) The same as step 1) of TEMFCC computation.
2) Next we calculate the Teager-energy of $s[\hat{n}]$:

$$\Psi(s[\hat{n}]) = s^2[\hat{n}] - s[\hat{n} - 1] s[\hat{n} + 1]$$

3) The magnitude spectrum of $\Psi(s[\hat{n}])$ is computed and warped to Mel-frequency scale, multiplied with a mel-filterbank(Fig. 2).

TABLE II
BERLIN EMOTIONAL SPEECH DATABASE

Emotion	Number of utterances
Anger	127
Happiness	71
Neutral	79
Sadness	62
Disgust	46
Fear	69
Boredom	69
Total	535

TABLE III
TESTING AND TRAINING SETS

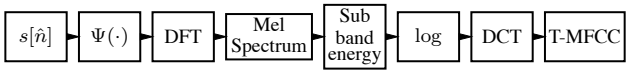
Male speakers		Female speakers
03, 10, 11, 12, 15		08, 09, 13, 14, 16
Experiment no.	Training set	Testing set
1	10, 11, 12, 15, 09, 13, 14, 16	03, 08
2	03, 11, 12, 15, 08, 13, 14, 16	10, 09
3	03, 10, 12, 15, 08, 09, 14, 16	11, 13
4	03, 10, 11, 15, 08, 09, 13, 16	12, 14
5	03, 10, 11, 12, 08, 09, 13, 14	15, 16

4) For each filterbank output of DFT$\{\Psi(s[\hat{n}])\}$, $\Psi_1[l]$, the DCT of the log of their magnitudes is computed to form:

$$\underset{k=1,..,N_c}{\text{T-MFCC}[\hat{n},k]} = \sum_{l=1}^{L} \log(\Psi_1[l]) \cos\left(\frac{k(l-0.5)}{L}\pi\right) \tag{19}$$

where T-MFCC$[\hat{n},k]$ is the k^{th} T-MFCC among N_c.

Figure below is a flowchart diagram describing the computation of T-MFCCs.



IV. EXPERIMENTAL FRAMEWORK

We explore the robustness and compare the performance of the proposed TEMFCC features to that of MFCCs and TEMFCCs by artificially injecting two types of noise to the speech signal and then computing their recognition accuracy.

The Berlin Emotional Speech Database (EMODB) [2], is used for simulation. EMODB is a recorded database of emotional utterances spoken by actors (i.e, simulated speech utterances). This database contains recordings, sampled at 16KHz, from 5 actors and 5 actresses, 10 different sentences of 7 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral are record (Table II). We have created the "EmoDB+Noise" by adding pink and white noise to the test set of EmoDB database where samples are distorted with white and pink noise respectively at SNR levels of 0, 10, 20, 30, 40, and 50dB.

We performed speaker-independent emotion recognition and the score reported for each feature is the average of five separate experiments. In each experiment, the speech utterances of a pair of speakers formed the testing set for the classifier while the remaining utterances formed the training set. The pairs were selected in order to include one male and female speaker at a time (Table III).

For every utterance endpoints are detected and the silent part was removed. Then the signal was divided into frames. The samples of each frame were weighted using hamming window to reduce the spectral leakage. The frame size used was 25.6 msec with 50% overlap between frames. Every speech frame is represented by three feature vectors:

- 12 MFCC coefficients
- 12 TEMFCC coefficients
- 12 T-MFCC coefficients

excluding the 0^{th} cepstrum coefficient c_0 and augmented with their 1^{st} and 2^{nd} time derivatives. The number of filters in the triangular Mel-filterbank used to extract the features vectors was 29.

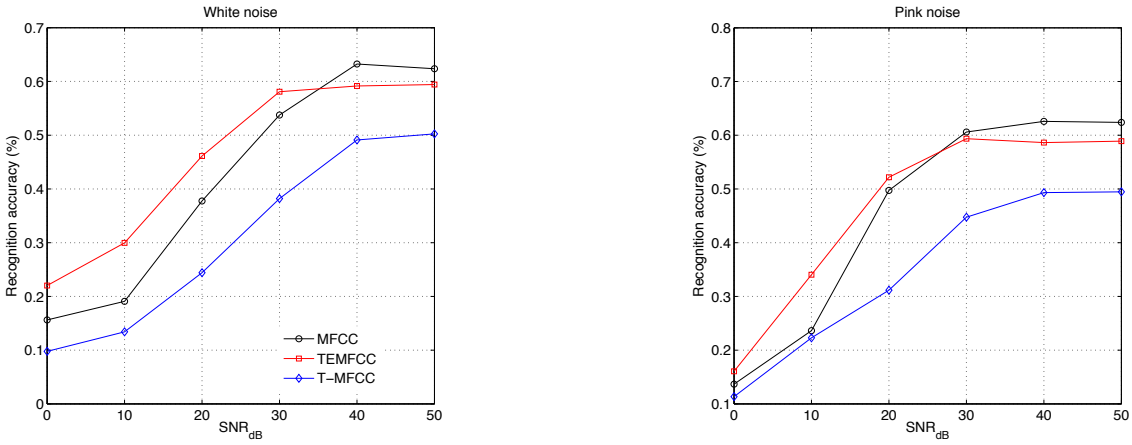
One of the key advantages of using differential parameters, such as delta cepstrum or delta-delta cepstrum, is that the differencing operation removes the effect of simple linear filtering on the parameter values, thereby making them less sensitive to channel shaping effects that might occur in a speech communication system. These frames are then classified into emotional states according to the maximum a posteriori probability (MAP) rule and the emotional class where the test signal belongs is the class where the majority of its frames belongs to.

Gaussian mixtures models (GMMs) were used to estimate the probability density function (pdf) of feature vectors in each emotional state. One problem we are faced when using GMMs for classification is how to choose the number of mixture components M. The CLUSTER software package has been used to automatically estimate model parameters from feature vectors representing speech frames. CLUSTER is an unsupervised algorithm for GMM estimation that is based on the expectation-maximization algorithm (EM) and the minimum description length (MDL) criterion [1].

Table IV presents the recognition rates for white and pink noise respectively. The greatest values are emphasized with bold fonts. In the case of white noise and SNR values less than or equal to 30dB, TEMFCCs have the best performance. As

TABLE IV
RECOGNITION ACCURACY

		SNR_{dB}					
	Noise	0	10	20	30	40	50
MFCC	White	0.16	0.19	0.38	0.54	**0.63**	**0.62**
	Pink	0.14	0.23	0.50	**0.60**	**0.62**	**0.62**
TEMFCC	White	**0.22**	**0.30**	**0.46**	**0.58**	0.59	0.59
	Pink	**0.16**	**0.34**	**0.52**	0.59	0.59	0.59
T-MFCC	White	0.10	0.13	0.24	0.38	0.49	0.50
	Pink	0.11	0.22	0.31	0.45	0.490	0.49



(a) White noise

(b) Pink noise

Fig. 3: Correct classification rate in the presence of (a) white and (b) pink noise for MFCC, TEMFCC, and T-MFCC

TABLE V
AVERAGE RECOGNITION RATES (%)

		Feature		
		MFCC	TEMFCC	T-MFCC
Noise	White	0.42	**0.46**	0.31
	Pink	0.45	**0.47**	0.34

the SNR values increase, MFCCs exhibit better performance. T-MFCCs have the overall lowest recognition rate in the case of white noise throughout all SNR levels. In the case of pink noise the results are quite similar, i.e., TEMFCCs have the best performance for values lower than 25dB, while MFCCs perform better than TEMFCCs and T-MFCCs as SNR values increase. As in the white-noise case, T-MFCCs have the overall lowest recognition rate in the range [0dB, 50dB] certifying that TEO in frequency domain provides better representations of non-linear variations of energy than in time domain. Figure 3 is a graphical representation of Table IV. Table V shows the average performance of MFCC, TEMFCC, and T-MFCC in presence of additive white and pink noise. It is observed that the average performance of TEMFCC features is better than MFCC, and T-MFCC features in additive noise environment.

V. CONCLUSIONS

In this paper we addressed the implementation of an automatic emotional-state recognition system capable of working in noisy environments, using cepstral features extracted from an audio signal. The proposed TEMFCC features have been shown to be more robust than MFCCs and T-MFCCs in white and pink noise environments for low SNR values. For clean conditions and white noise the TEMFCCs performed similarly to the MFCCs. T-MFCCs have the lowest recognition accuracy. The experiments were carried out using the EMODB speech corpus. The increased robustness of TEMFCCs is due to fact that TEO performs a demodulation-like operation and the envelope of the spectrum produces more robust features, but the physical interpretation of applying the TEO after the Fourier transform should be further investigated.

REFERENCES

[1] C. A. Bouman. Cluster: An unsupervised algorithm for modeling Gaussian mixtures. 1997. Available from http://www.ece.purdue.edu/~bouman.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Proceedings of Interspeech, Lissabon*, pages 1517–1520, 2005.

[3] Dimitrios Dimitriadis, Alexandros Potamianos, and Petros Maragos. A comparison of the squared energy and teager-kaiser operators for short-term energy estimation in additive noise. *Trans. Sig. Proc.*, 57(7):2569–2581, July 2009.

[4] Dimitriadis Dimitrios, Maragos Petros, and Potamianos Alexandros. Auditory teager energy cepstrum coefficients for robust speech recognition. *INTERSPEECH*, 2005.

[5] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, 1983.

[6] Petros Maragos and Alan C. Bovik. Image demodulation using multi-dimensional energy separation. *J. Opt. Soc. Am. A*, 12(9):1867–1876, 1995.

[7] N.S. Nehe and R.S. Holambe. Power spectrum difference teager energy features for speech recognition in noisy environment. In *Industrial and Information Systems, 2008. ICIIS 2008. IEEE Region 10 and the Third international Conference on*, pages 1–5, Dec 2008.

[8] T. L. Nwe, S. W. Foo, and L. C. De Silva. Detection of stress and emotion in speech using traditional and FFT based log energy features. *Proceedings of the 4th International Conference on Information, Communications and Signal Processing*, 2009.

[9] Hemant A. Patil and T. K. Basu. Identifying perceptually similar languages using teager energy based cepstrum. *Engineering Letters*, 16(1):151–159, 2008.

[10] Hemant A. Patil and Keshab K. Parhi. Novel variable length teager energy based features for person recognition from their hum. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4526 –4529, March 2010.

[11] Thomas F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall Signal Processing Series. Prentice Hall, 2001.